Machine Learning in Diagnostic Imaging - Methodologic Considerations

Constantine Gatsonis, PhD
Department of Biostatistics and
Center for Statistical Sciences
Brown University School of Public Health

June 2020

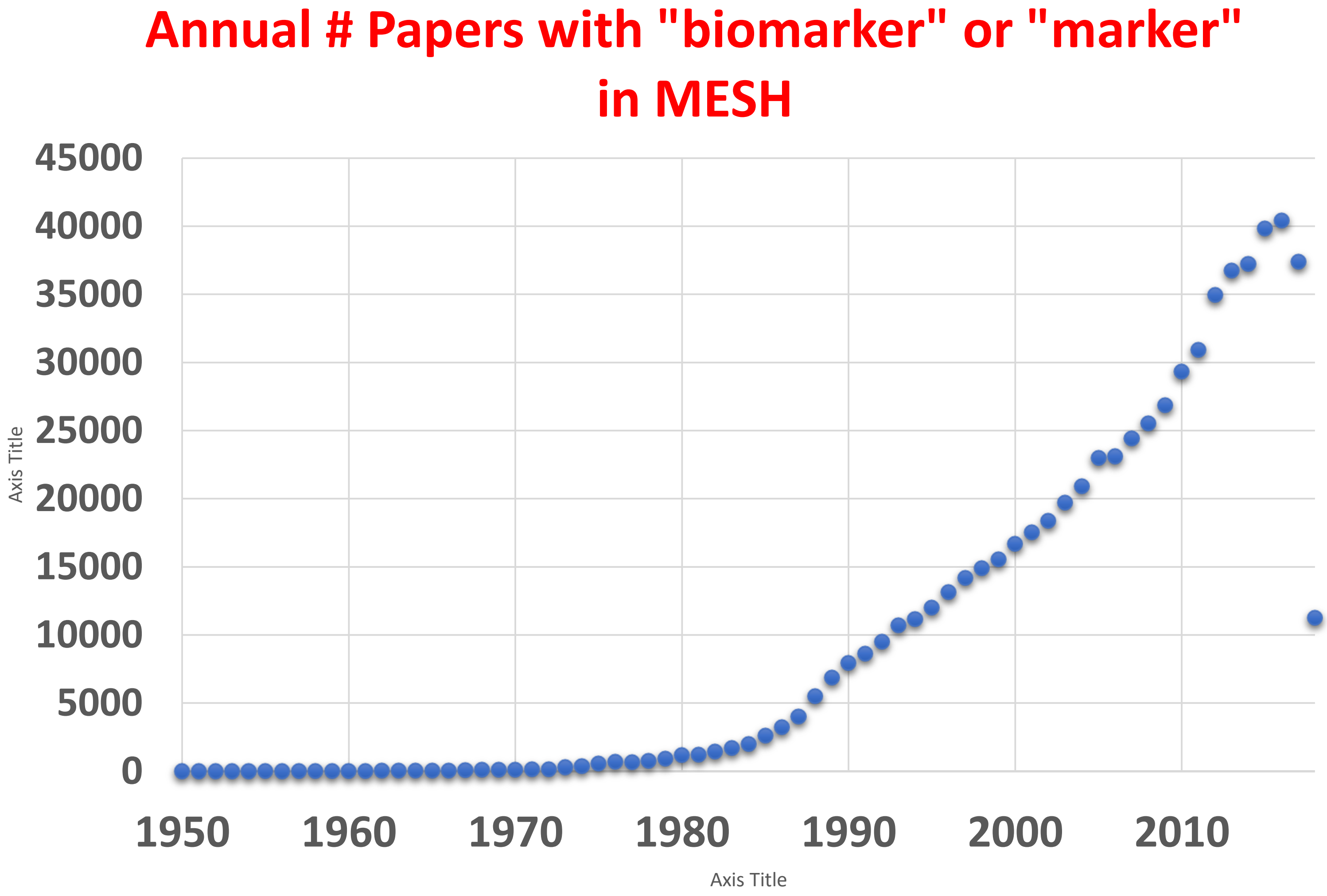
Purdue University

BROWN
School of Public Health

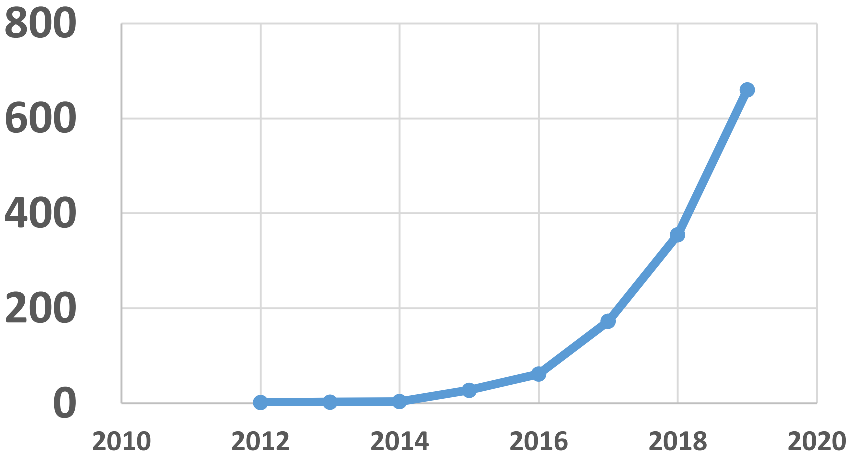
Outline

- *Radiomics* uses statistical machine learning methods to derive knowledge from medical images.
- The discovery space for radiomics-based markers has grown impressively.
- However, substantial challenges arise in the translational space.
- Focus on radiomics-based markers for clinical care and clinical trials.
- Emphasis on markers based on Deep Learning methods.

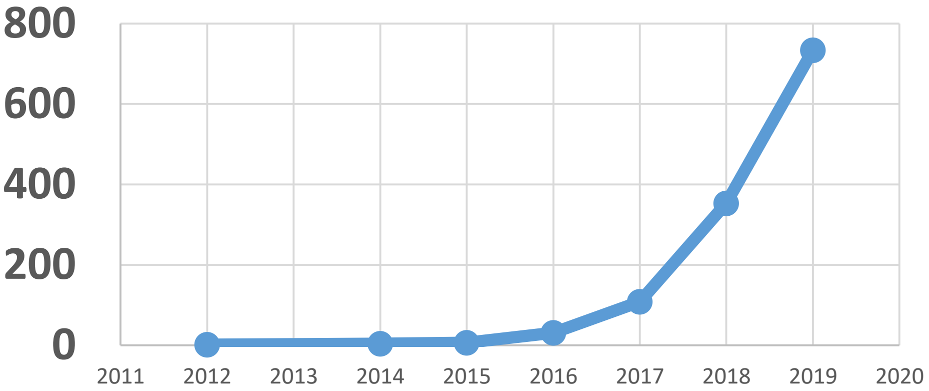
The growth in biomarker research continues



Articles with "radiomics" in title/abstract

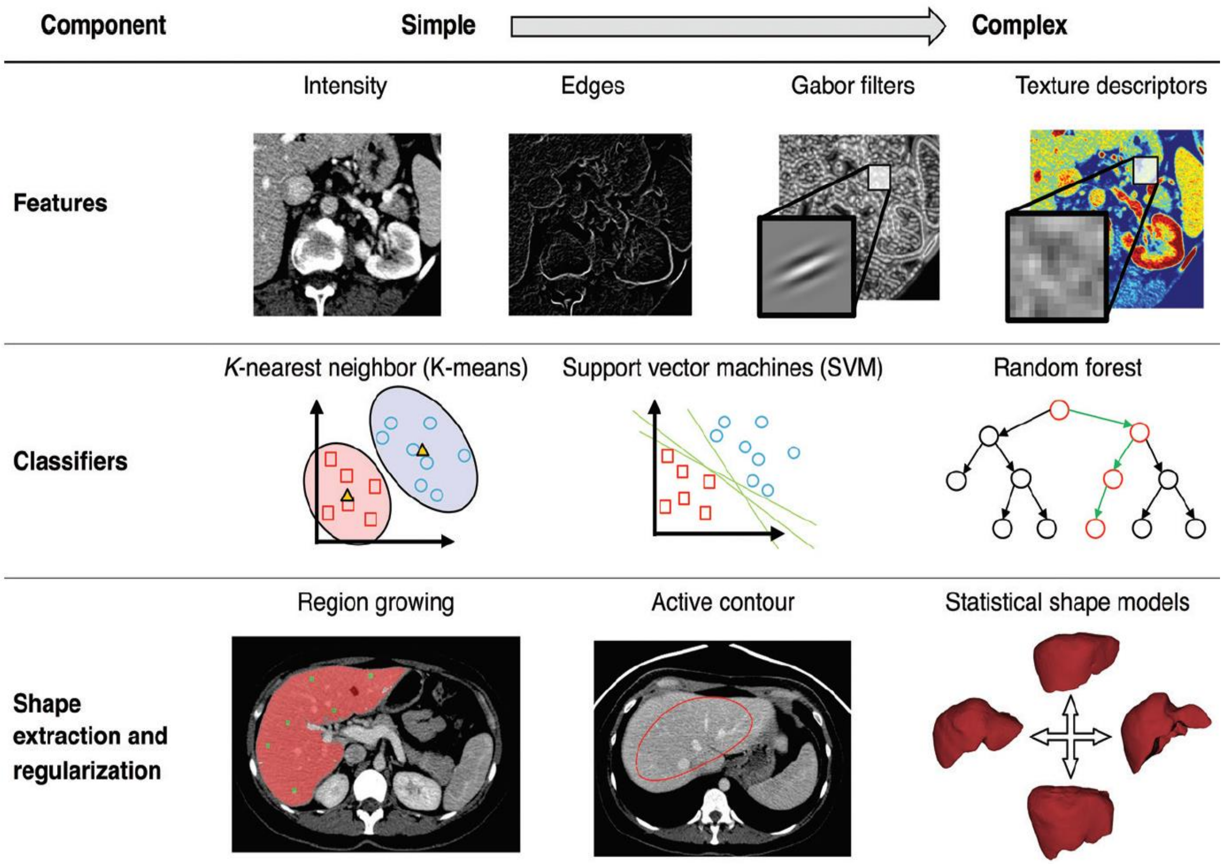


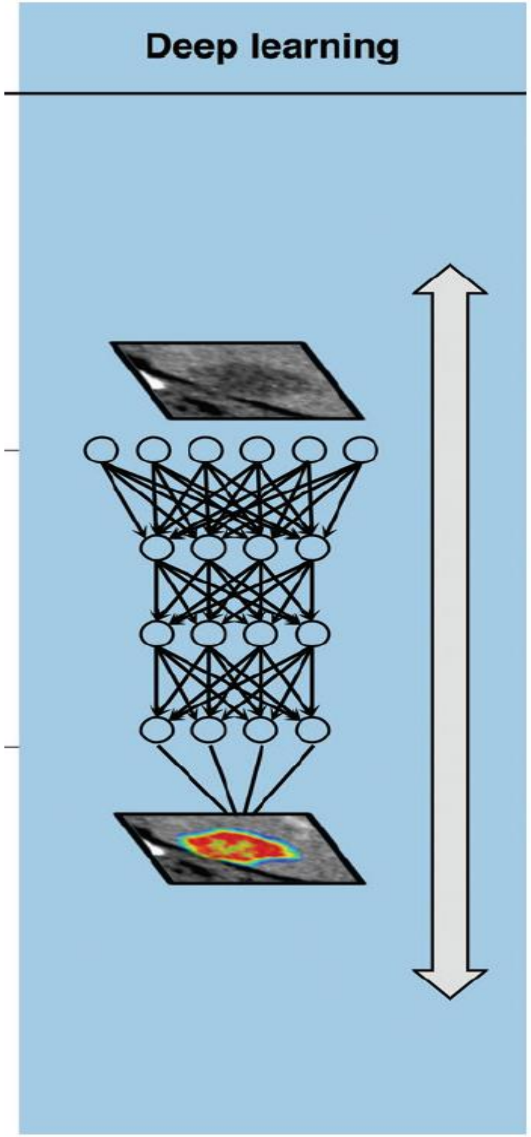
Articles with "deep learning" and "diagnosis" or "imaging" in title/absrract



Spectrum of radiomics methods

Feature space analysis

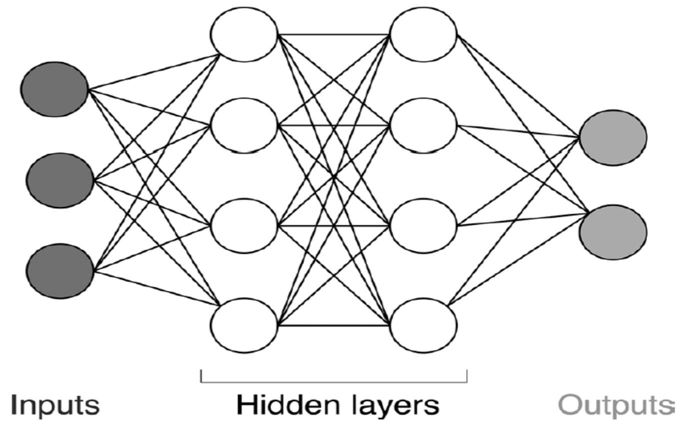




From : Chartrand et al, Radiographics 2017

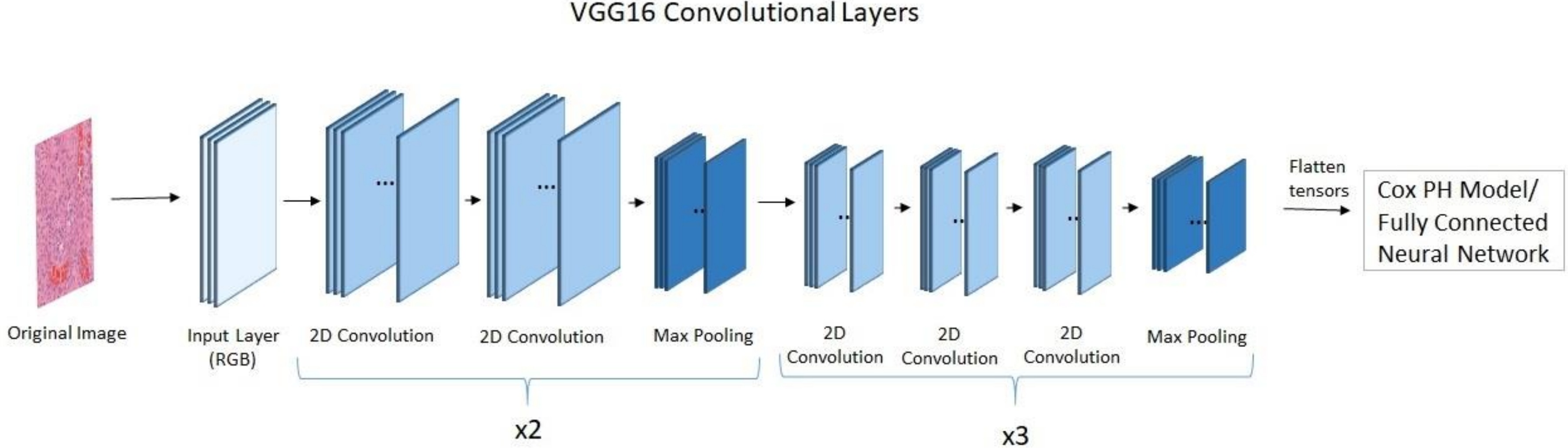
Architecture of multi-layer NNs

Common deep learning network



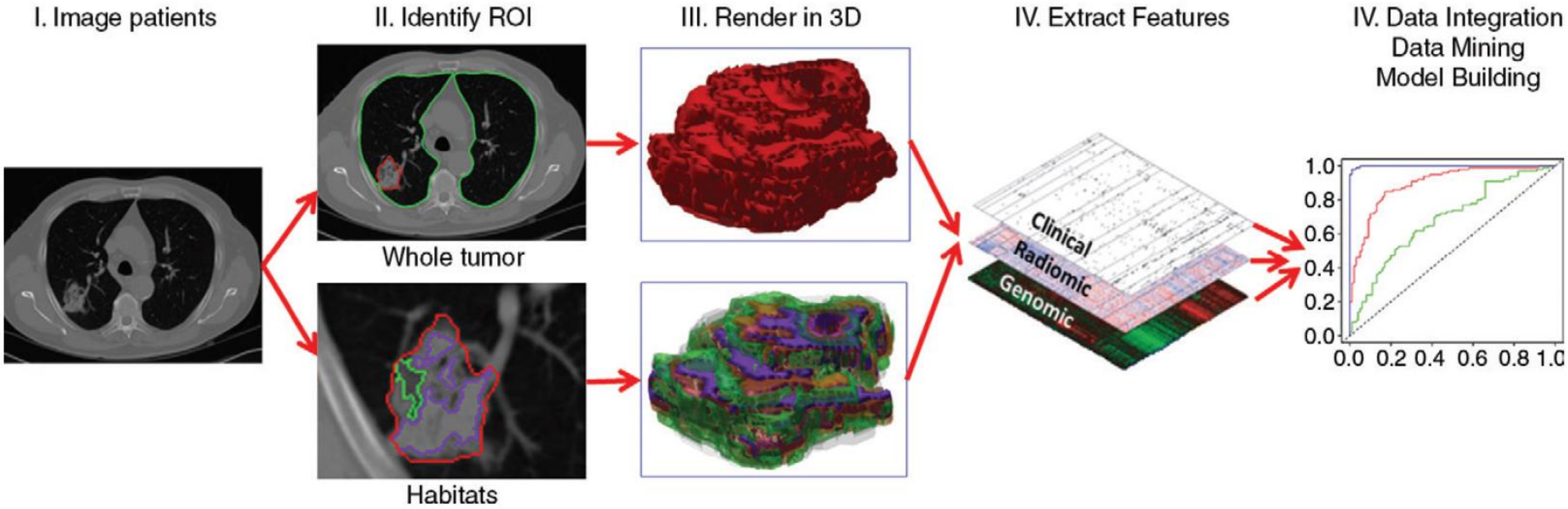
From : Chartrand et al,

Convolution NNs (CNN) used in analysis of imagine



From : S. Morrison et al, work in progress,

Radiomics in high dimensional feature space

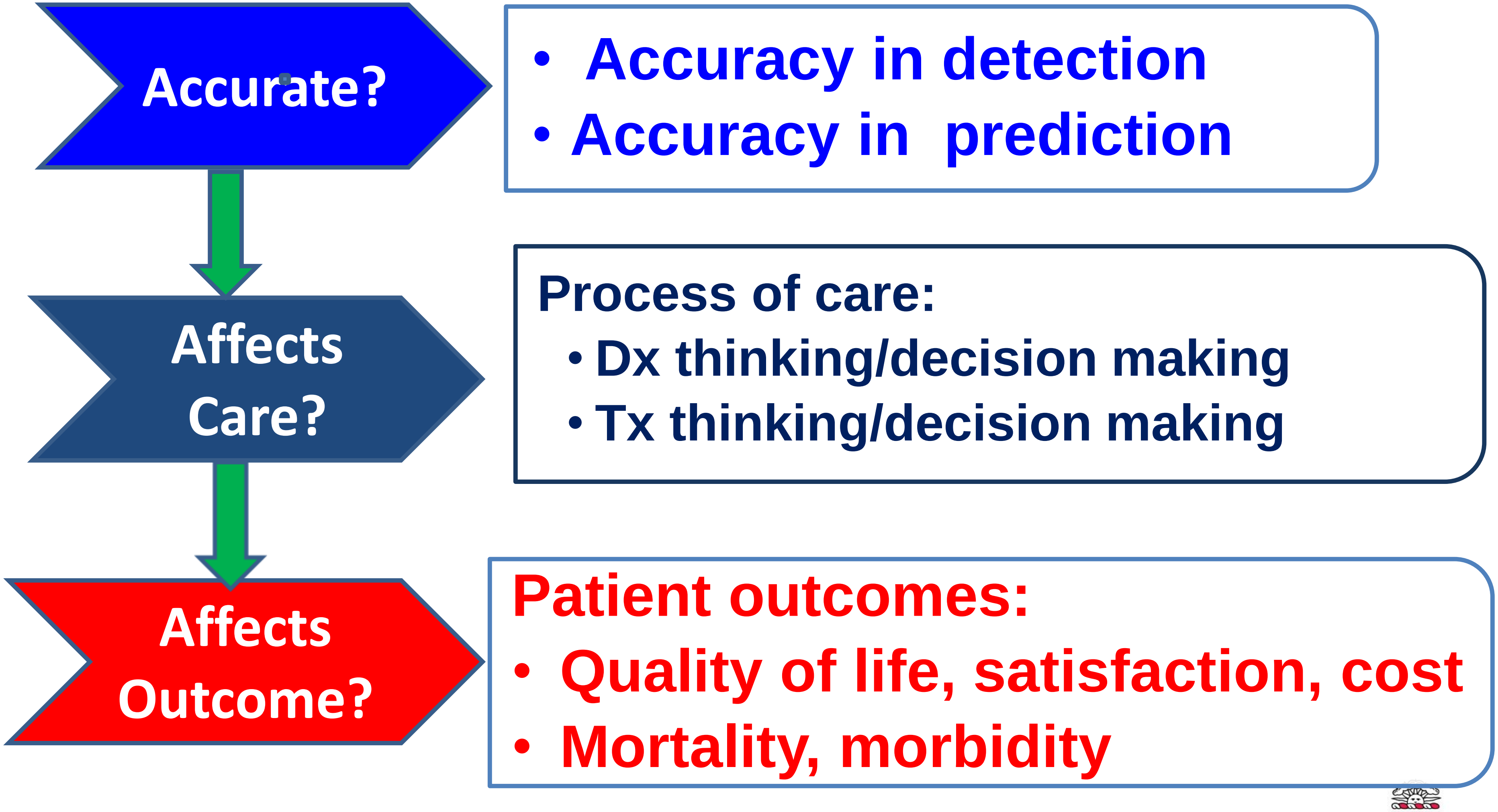


Gillies, Radiology 2016

Key aspects

- Segmentation
- Feature definition and extraction
  - Semantic features (e.g. shape, vascularity, necrosis)
  - Agnostic features (e.g. histogram of signal intensity, various transforms)
- Classifier modeling

Evaluating radiomic markers in the clinical setting



Accurate?

- **Accuracy in detection**
- **Accuracy in prediction**

Affects Care?

Process of care:

- **Dx thinking/decision making**
- **Tx thinking/decision making**

Affects Outcome?

Patient outcomes:

- **Quality of life, satisfaction, cost**
- **Mortality, morbidity**

Schematic of evolution and evaluation of markers

Stage I: Discovery.

Present status for most radiomics markers

Stage II : Introductory

Typically single institution studies

Stage III: Mature

Multi-institutional studies

Stage IV: Disseminated

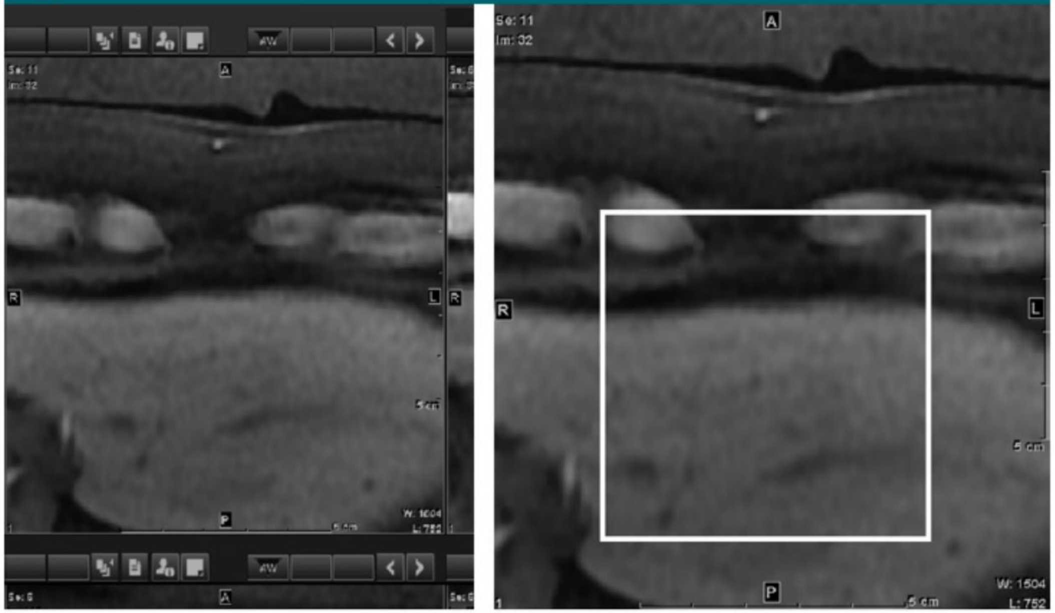
Observational studies, registries

Some recent examples of deep learning studies

Deep learning: A recent example

Liver Fibrosis: Deep Convolutional Neural Network for Staging by Using Gadoxetic Acid–enhanced Hepatobiliary Phase MR Images. Yasaka et al, Radiology, April 2018

Figure 1



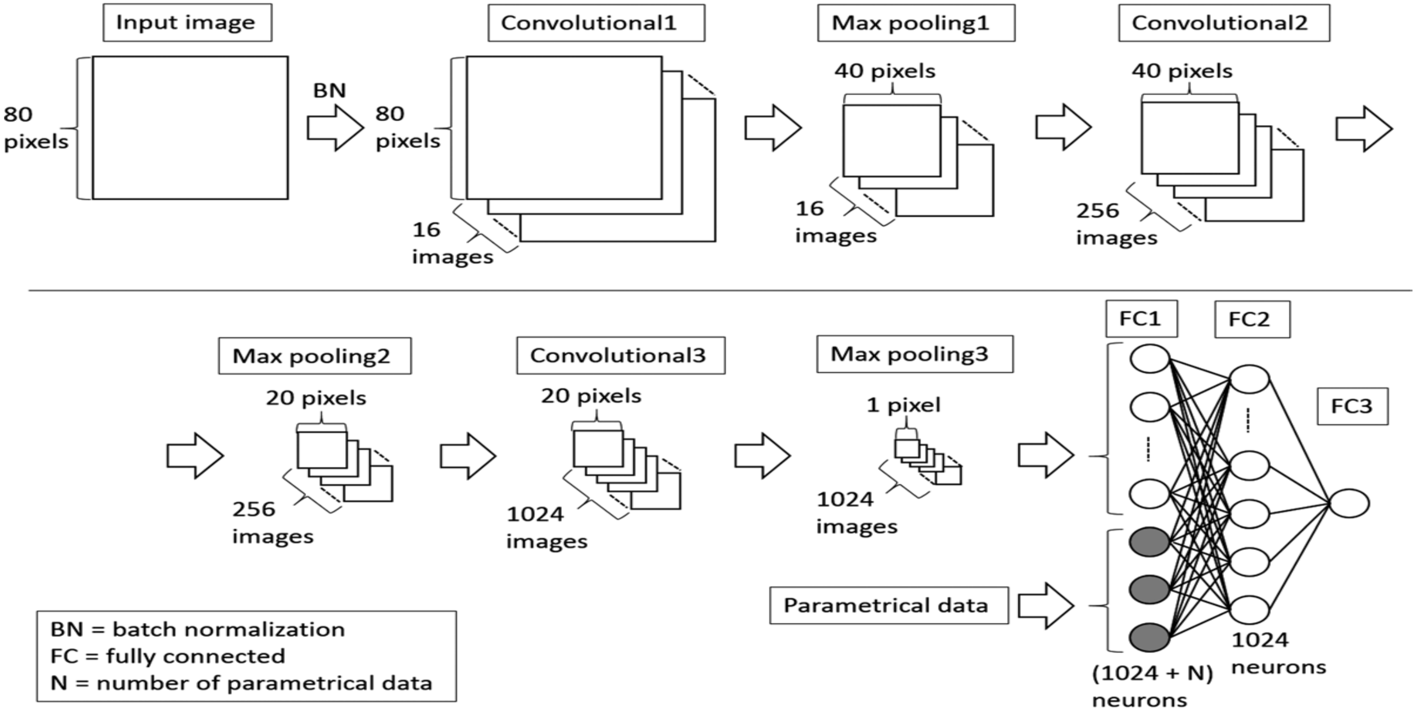
a. b.

Figure 1: Image data format process. **(a)** The images were magnified on a commercial viewer, referencing the scale bar shown at the bottom of the window. **(b)** The captured images (594 × 644 pixels) were cropped with a square crop box (white square) (350 × 350 pixels). The cropped images (350 × 350 pixels) were resized to 80 × 80 pixels before they were fed to the DCNN.

Training set: 534 patients

Test set: 100 patients

MRI: 1.5T and 3T



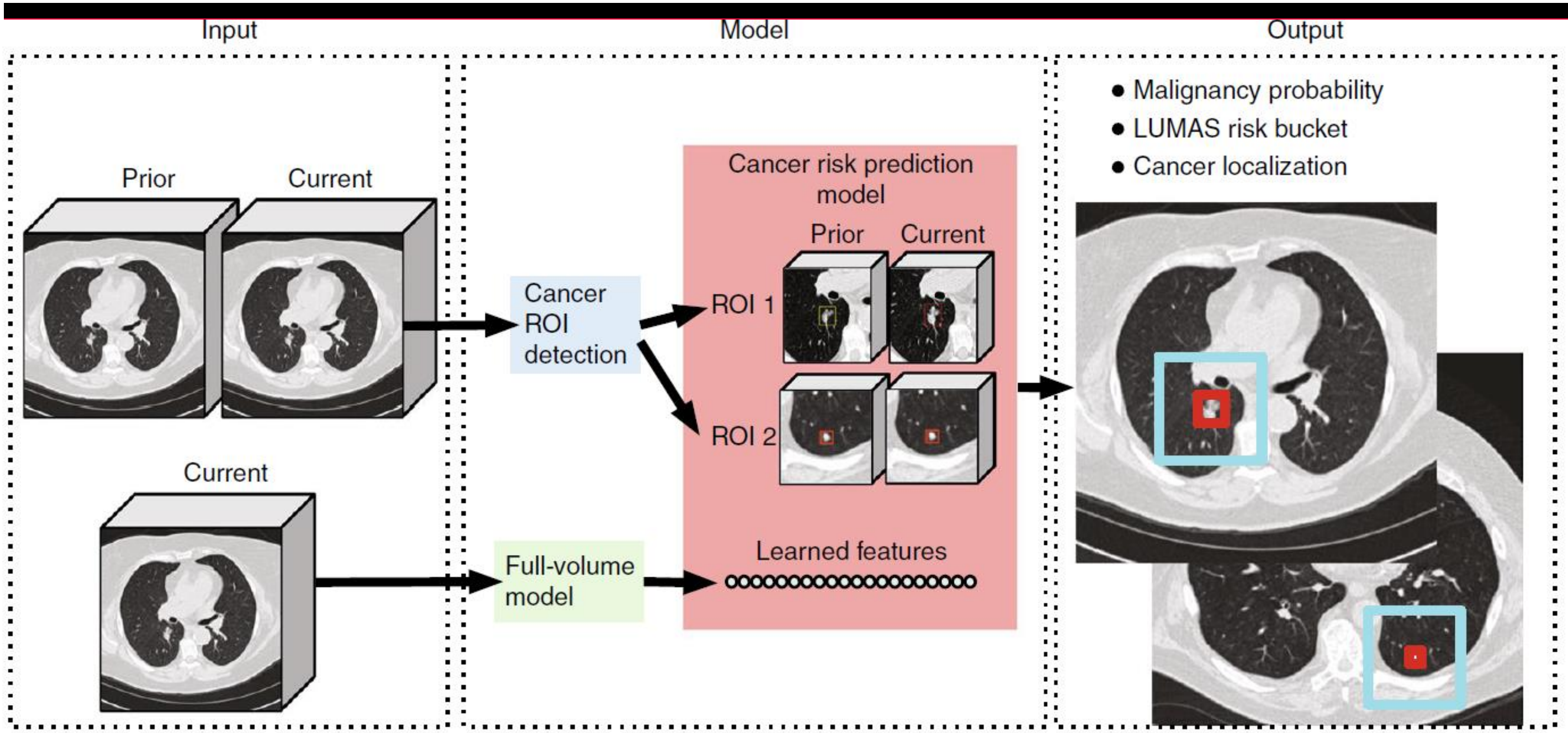
Schematic of DCNN in Liver Fibrosis analysis, From Yasaka et al

- **MRI images in training session were augmented (90 augmented images per original)**
- **CNN included information on HBV and HCV status**
- **Supervised training**
- **Fibrosis score F_{DL} was derived.**

Diagnostic Performance of the F_{DL} score for staging liver fibrosis in the Test Data Set. From Yasaka et al

	Cirrhosis (F4 vs F3–0)	Advanced Fibrosis (F4–3 vs F2–0)	Substantial Fibrosis (F4–2 vs F1–0)
Full model			
AUC	0.84 (0.81–0.85)	0.84 (0.83–0.86)	0.85 (0.82–0.86)
Threshold	3.37 (3.31–3.52)	2.89 (2.79–3.03)	2.22 (2.11–2.49)
Sensitivity	0.76 (0.72–0.79)	0.78 (0.75–0.85)	0.84 (0.83–0.86)
Specificity	0.76 (0.74–0.77)	0.74 (0.70–0.77)	0.65 (0.60–0.68)

DL for lung cancer screening



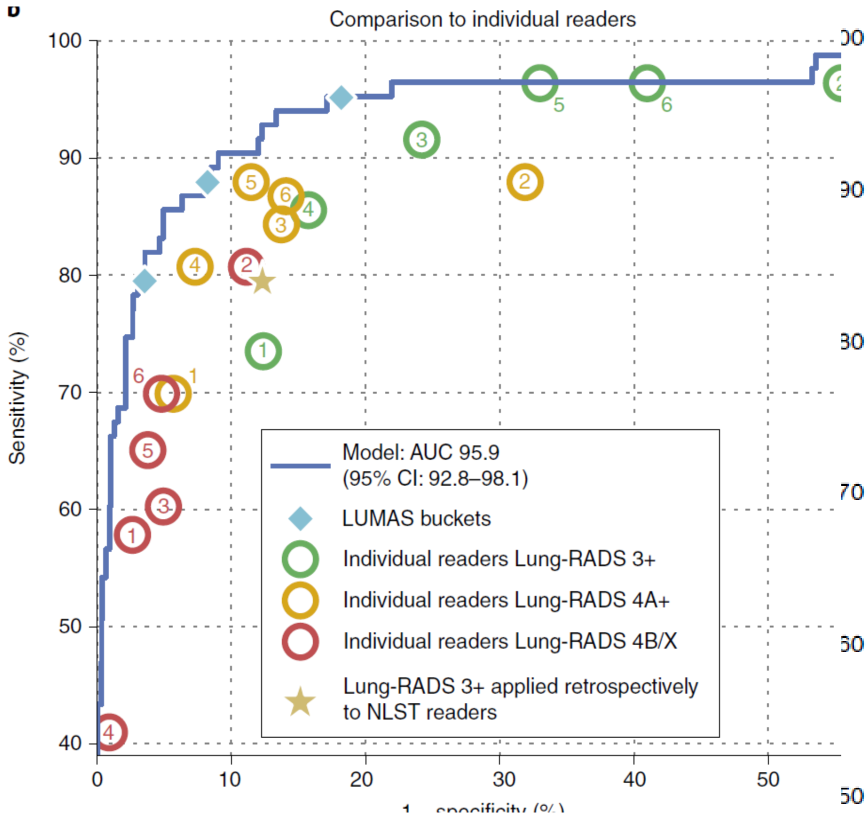
End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography

Ardila et al, Nature Medicine 2019

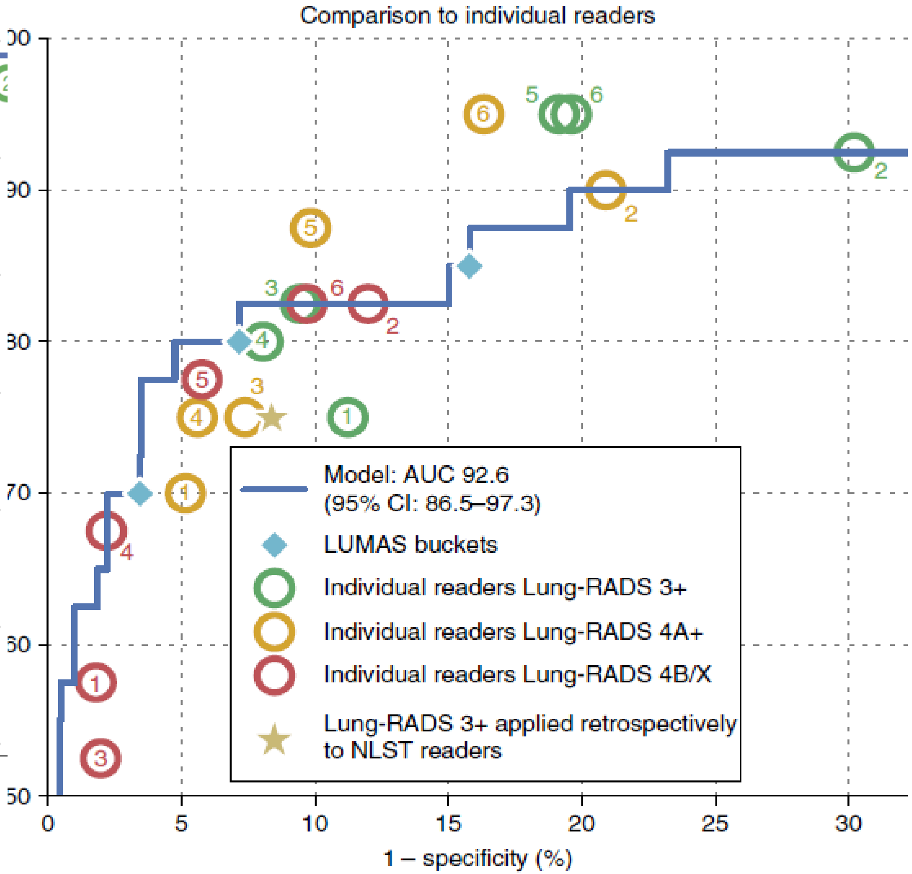
DL analysis of images from the National Lung Screening Trial (NLST)

Subset of 6717 cases.

Prediction of malignancy of model vs human interpreters



Using current CT



Using current and prior CT

…This creates an opportunity to optimize the screening process via computer assistance and automation. While the vast majority of patients remain unscreened, we show the potential for deep learning models to increase the accuracy, consistency and adoption of lung cancer screening worldwide

Deep learning prediction of time-to-event response

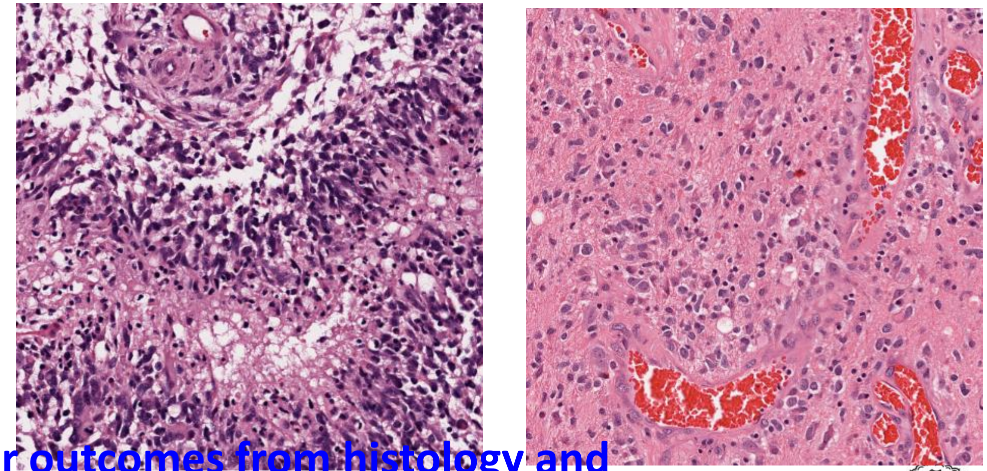
Samantha Morrison, Jon Steingrimmson, CG

Work in progress

PLEASE DO NOT QUOTE WITHOUT PERMISSION

- Brain cancer histology from TCGA
- H&E stained whole slide had ROIs identified by experts.
- These regions of interest were magnified (20x) and used as inputs to the modeling process (1024 x 1024 pixels)

Histology ROIs from two participants.
Survival times:
627 days and 1077 days.



P. Mobadersany, et al. Predicting cancer outcomes from histology and genomics using convolutional networks. PNASciences, 2018.

Extracting features from images

- Analysis uses a pre-trained network
  - ImageNetVGG16 (Oxford Visual Geometry Group)
- Input: 1024 x 1024 pixel images
- Output for each image: tensor 32 x 32 x 512
- Output tensors used as input in further analysis, e.g.
  - Regularized Cox regression modeling
  - Densely connected neural network
- Approach reduces time and computational burden

Extracting features from images

- Analysis uses a pre-trained network
  - ImageNetVGG16 (Oxford Visual Geometry Group)
- Input: 1024 x 1024 pixel images
- Output for each image: tensor 32 x 32 x 512
- Output tensors used as input in further analysis, e.g.
  - Regularized Cox regression modeling
  - Densely connected neural network
- Approach reduces time and computational burden

Work in progress. Please do not quote without permission

VGG16 analysis – cont'd

- VGG16
  - Improves classification accuracy by increasing depth of neural network with small convolutional layers (3 x 3)
  - Small convolutional layers decreases computation burden and number of parameters
  - VGG16 CNN was trained on variety of augmented images.
- In part of the analysis we removed the last 3 densely connected layers, keeping only convolutional layers.
- Convolution layers include: 2D convolutions, Max Pooling, and ReLU activation function

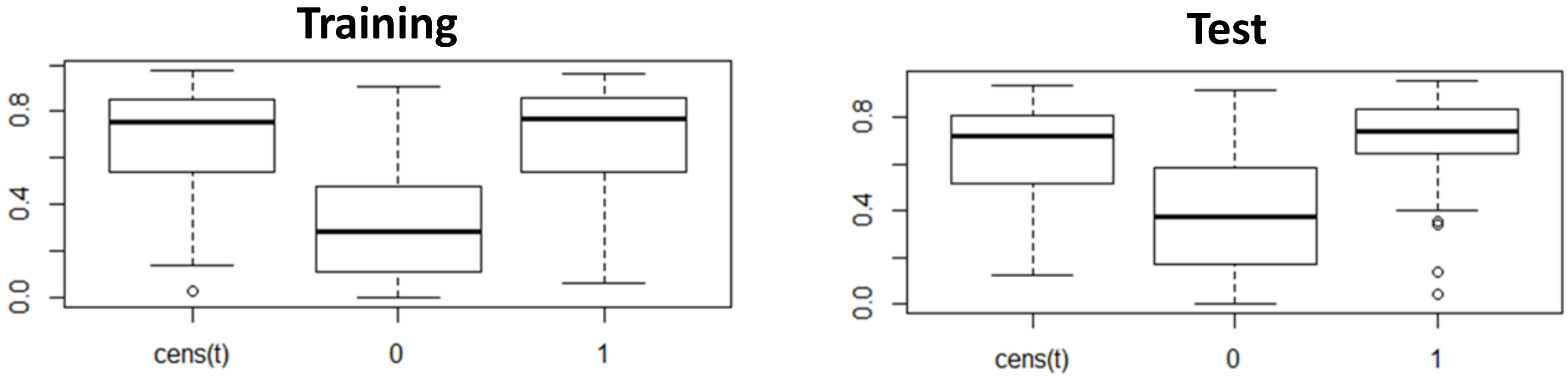
K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014

VGG16 + Cox Regression

Input:

1024x1024 pixel images → **Pre-trained vgg16 (convolutional layers)** → **vgg16 features (each 32 x 32 x 512)** → Flatten features → **Cox regression with lasso (28 features)**

Prediction of survival >914 days



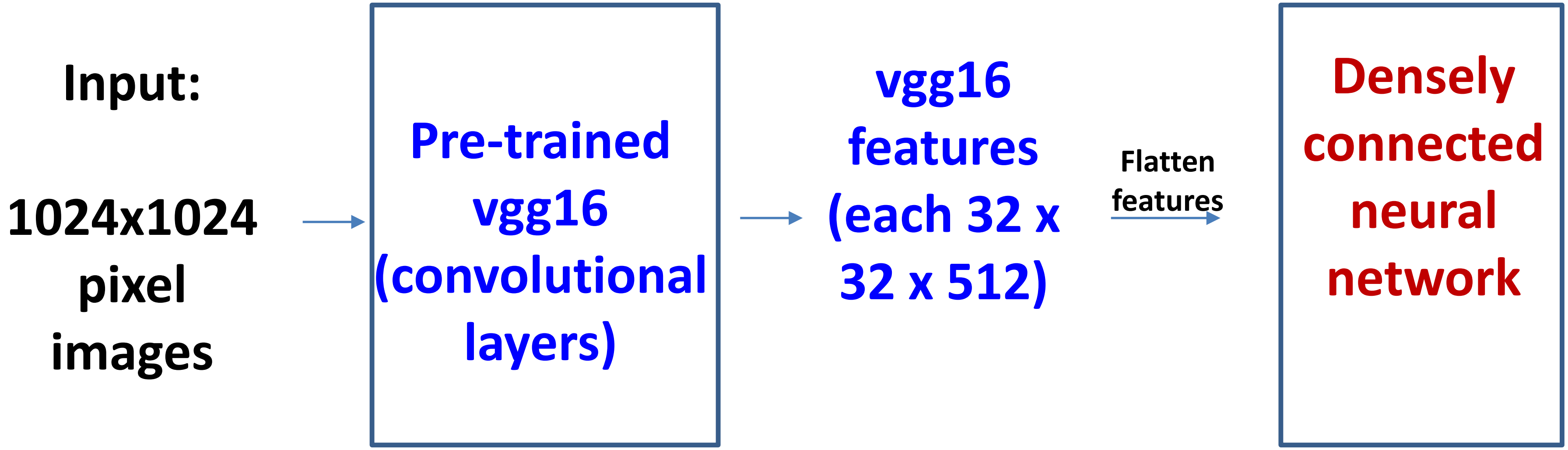
Training

Test

Work in progress. Please do not quote without permission

BROWN
School of Public Health

VGG16 + FCN



Input:

1024x1024 pixel images

Pre-trained vgg16 (convolutional layers)

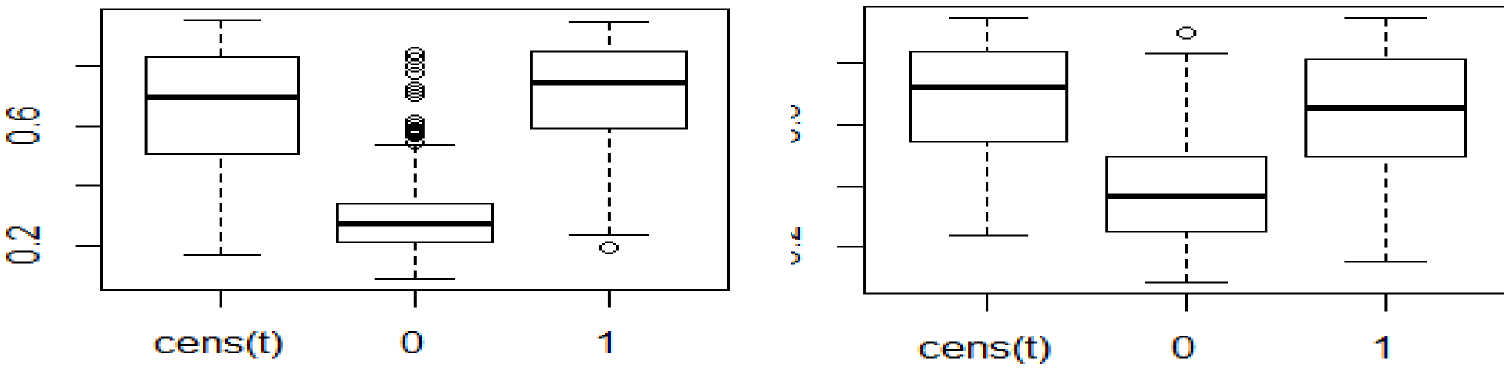
vgg16 features (each 32 x 32 x 512)

Flatten features

Densely connected neural network

Prediction of survival >914 days

Training **Test**

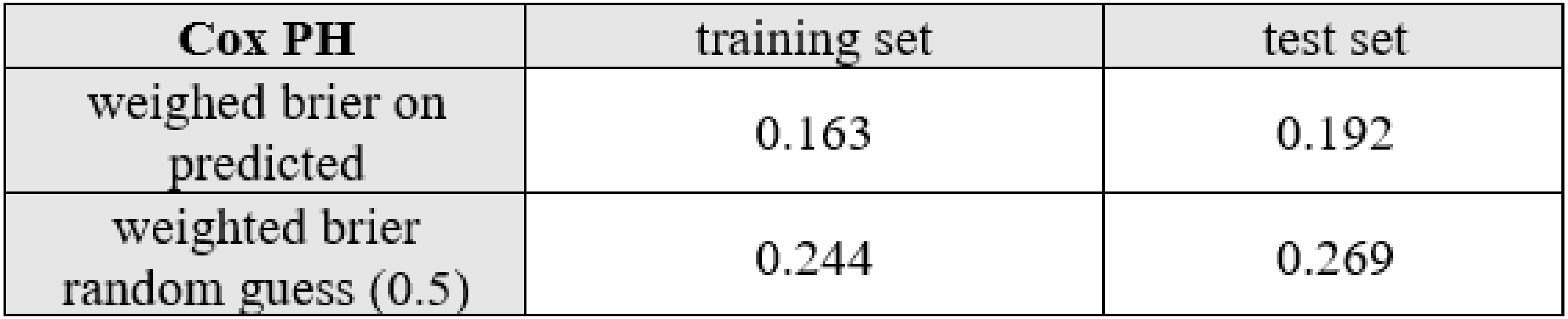


Work in progress. Please do not quote without permission

BROWN
School of Public Health

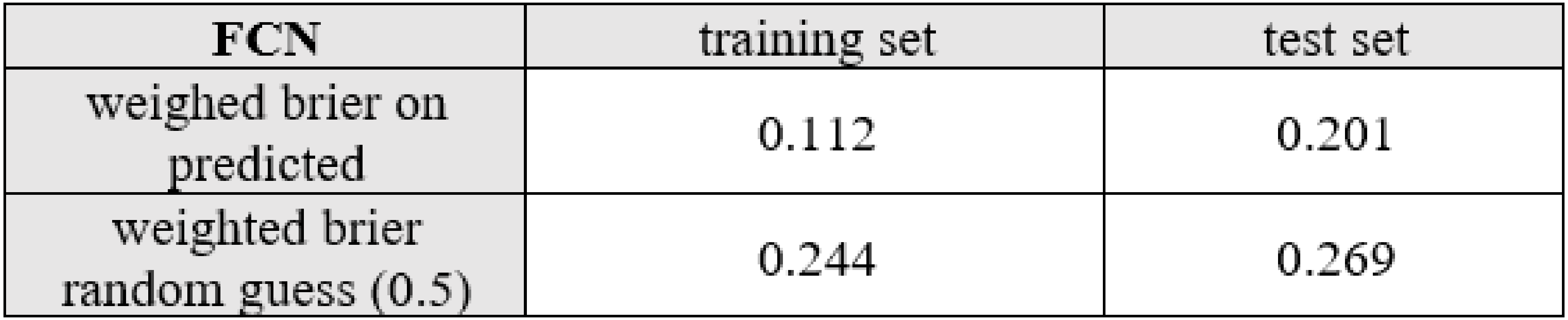
Weighted Brier Scores

Cox regression



Cox PH	training set	test set
weighed brier on predicted	0.163	0.192
weighted brier random guess (0.5)	0.244	0.269

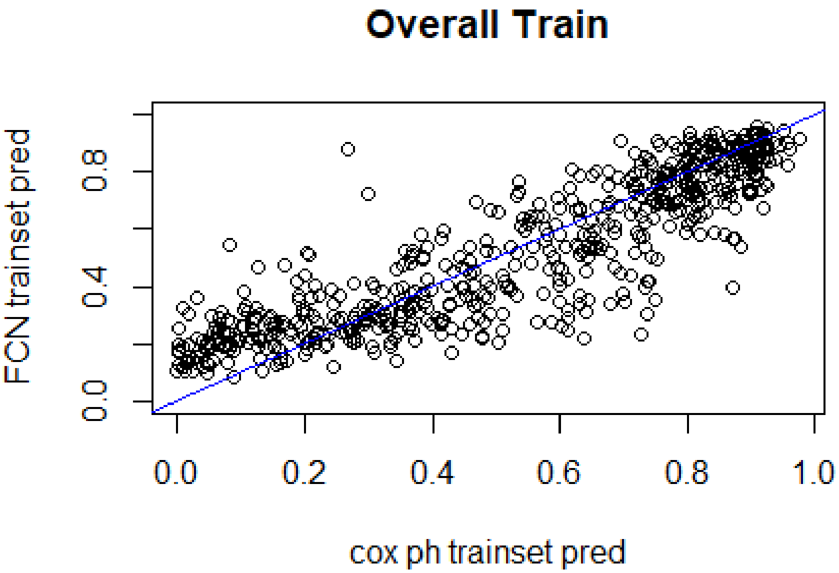
FCN

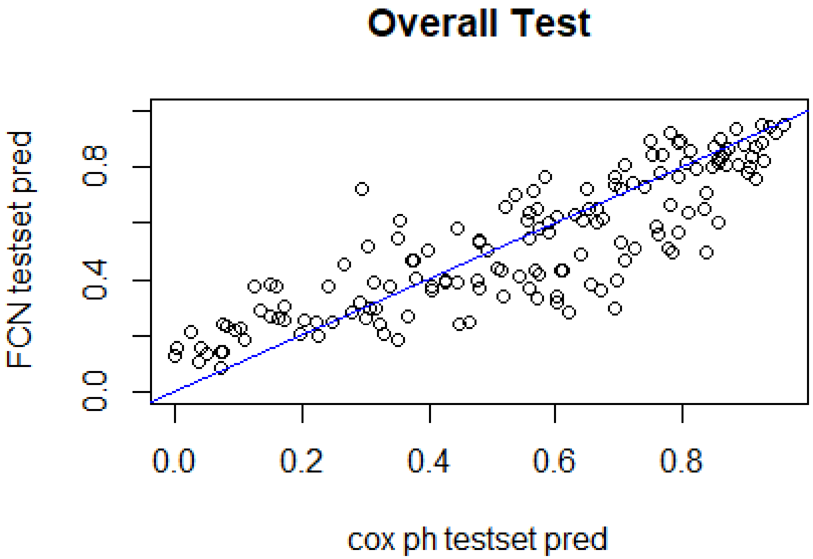


FCN	training set	test set
weighed brier on predicted	0.112	0.201
weighted brier random guess (0.5)	0.244	0.269

Work in progress. Please do not quote without permission

Comparison of predictions

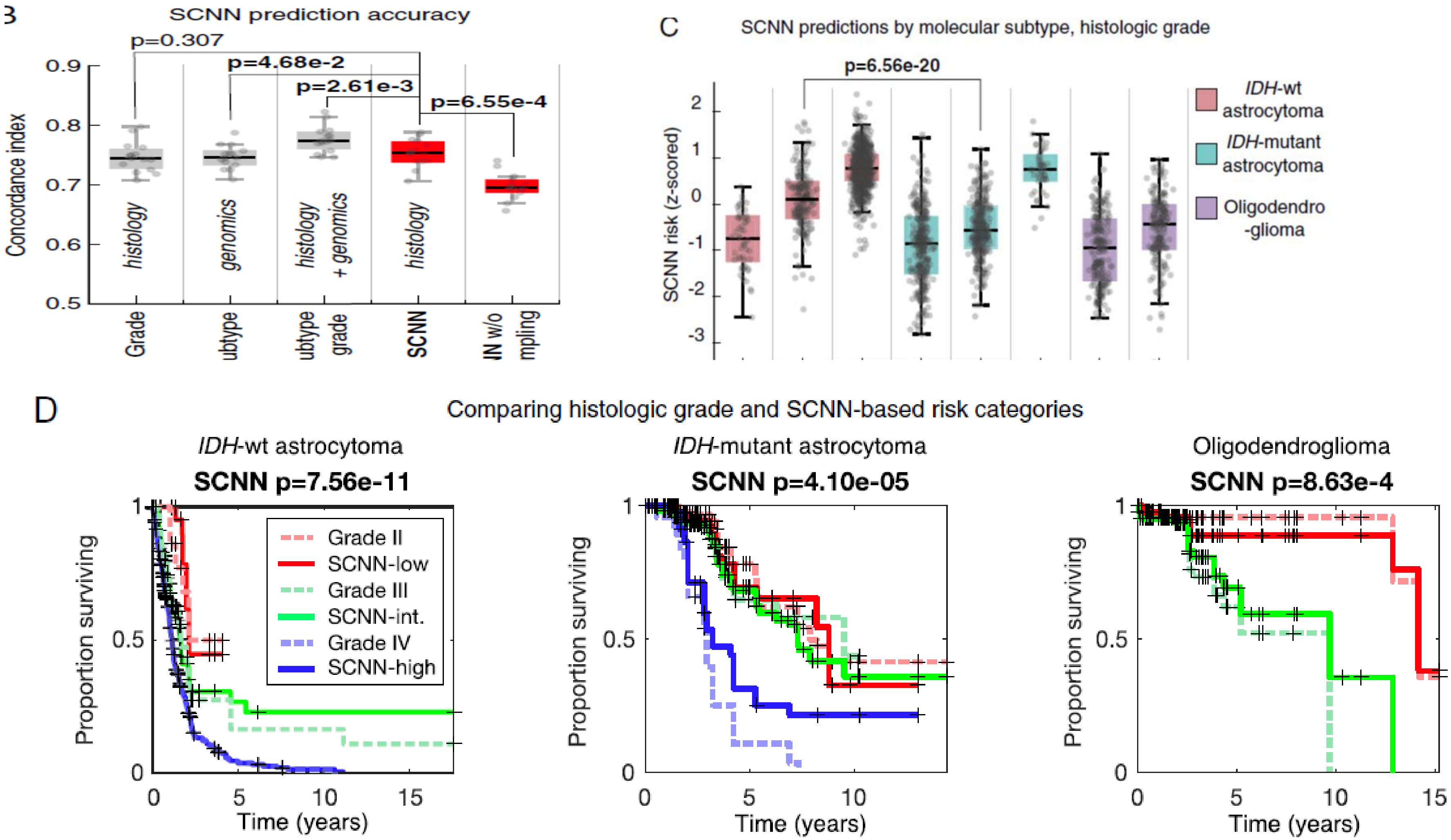




Work in progress. Please do not quote without permission

Radiogenomics analysis

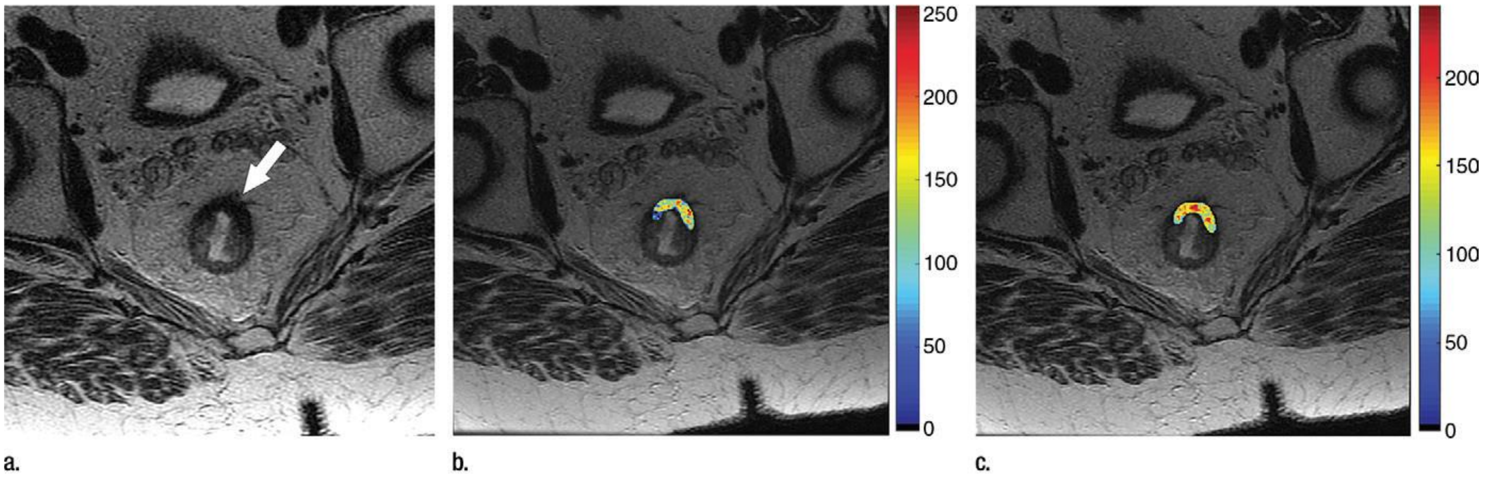
From Mobadersany et al, PNAS 2018

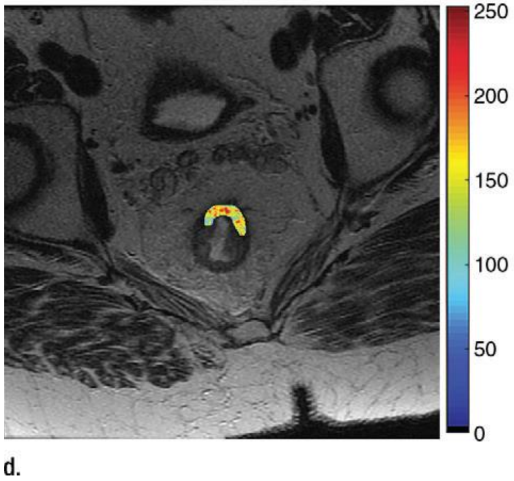


Some recent examples of feature-based (high dimensional) analysis

MR Imaging of Rectal Cancer: Radiomics Analysis to Assess Treatment Response after Neoadjuvant Therapy

Horvat et al Radiology 2018





- 141 patients
- 21 had pCR, 93 had PR
- T2-weighted MRI features radiomics
- T2- and DW weighted qualitative assessment
- 34 features computed
- Random Forest classifier

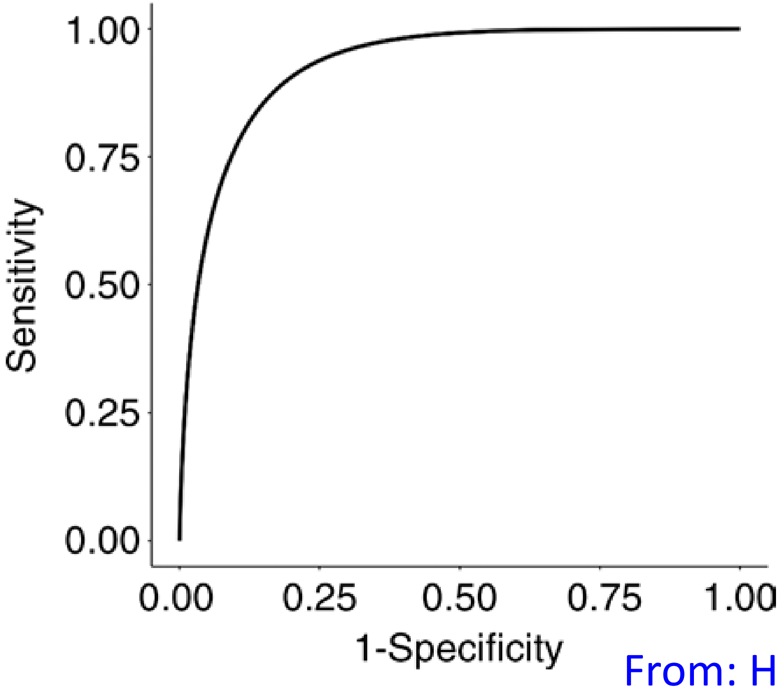
Radiomic features and their performance

Feature	Gini Importance	Median pCR	Median pPR	P Value
Energy	0.99	84.5	68.1	0.005
Kurtosis	0.95	3.7	4.9	0.04
Homogeneity	0.82	71.6	50.2	0.005
Gab45.contrast	0.78	63.6	95.7	0.003
Gab45.entropy	0.69	104.8	128.3	0.006
Gab90.contrast	0.66	70.9	93.9	0.006
Contrast	0.61	18.8	9.9	0.001
Gab0.entropy	0.58	105.5	130.1	0.006

Excerpt of table from: Horvat et al Radiology 2018

Diagnostic and predictive performance of radiomic index for pCR

Sensitivity	100 (84, 100)
Specificity	91 (84, 96)
PPV	72 (53, 87)
NPV	100 (96, 100)



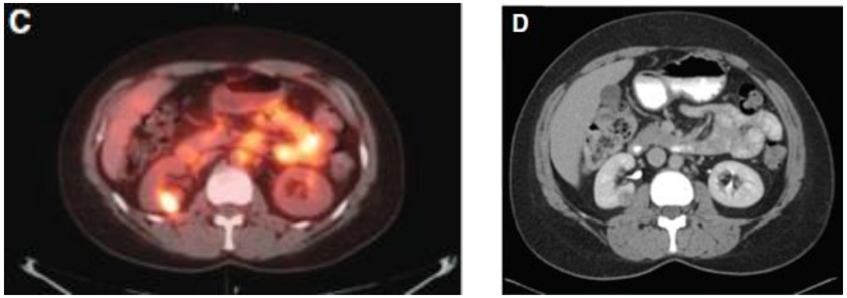
From: Horvat et al Radiology 2018

Radiomic analysis for REDECT study

Ongoing project
Brown: Samantha Morrison, CG
Columbia: F. Ahmed, L. Liu, B. Zhao

Original trial conducted to assess the performance of Iodine-124-girentuximab PET/CT in the detection of clear cell carcinoma (ccRCC) in patients with renal cancer.

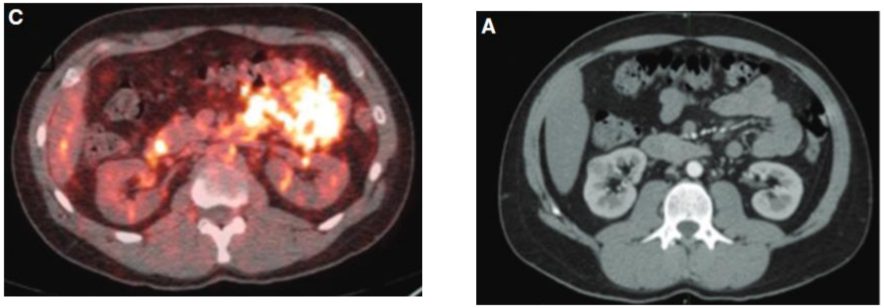
Divgi CG et al.,JCO 2013



C) Iodine-124-girentuximab PET/CT fused image

D) Contrast enhanced CT (CECT)

Pathology: 1.0 cm right renal clear cell carcinoma



A) Contrast enhanced CT (CECT)

C) iodine-124-girentuximab PET/CT scan

Pathology: 1.8 cm right renal oncocytoma

Radiomic features extracted

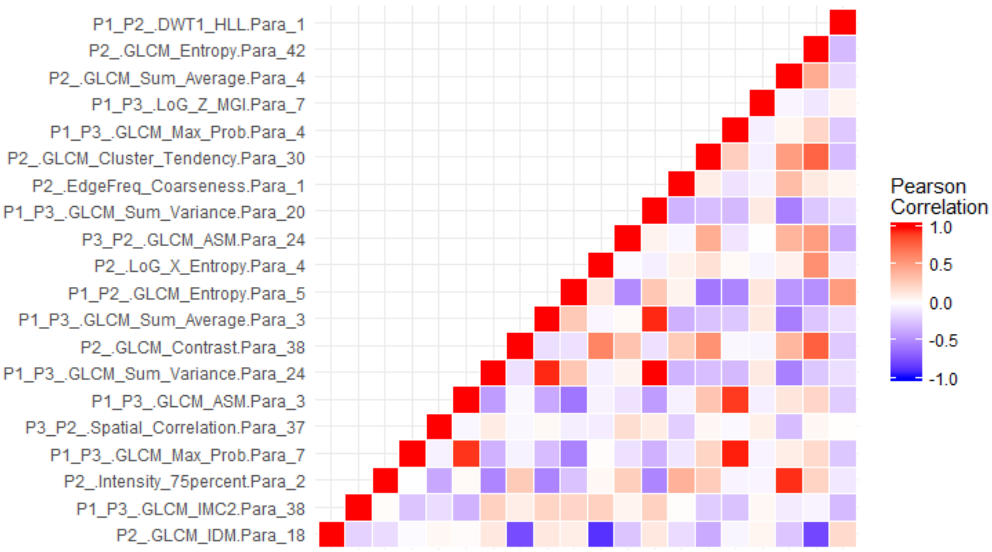
190 cases, 5287 features extracted from each case

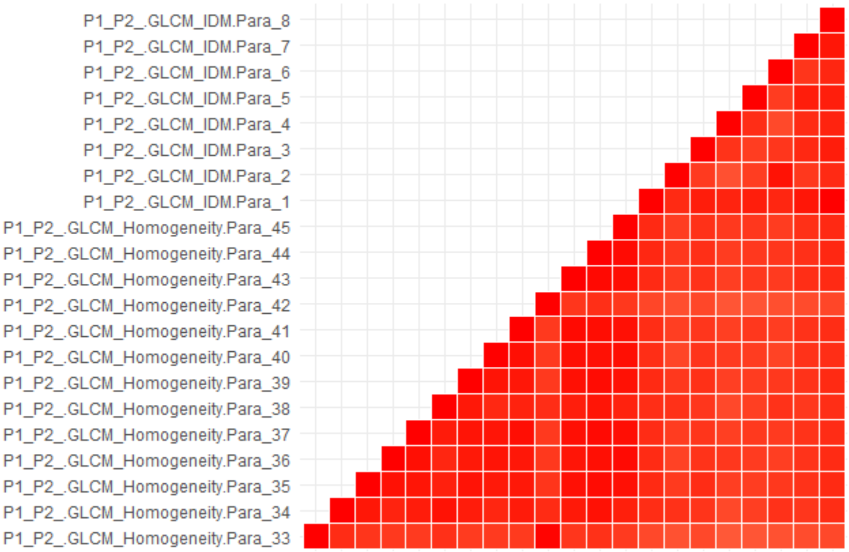
Groups of features

- Size Related
- First order statistics
- Shape
- Surface Shape
- Sigmoid Functions
- Wavelet Features (DWT, DWF)
- Edge Frequency features
- Fractal Dimension
- GTDM (Gray Tone Difference Matrix)
- Gabor Energy
- Laws' Energy
- Laplacian of Gaussian (LoG)
- Run Length features
- Spatial Correlation
- GLCM (Gray Level Co occurrence Matrix)

Correlation in features- examples

High correlation among features



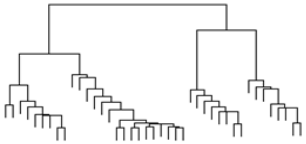


Data reduction and model fitting

5287 features

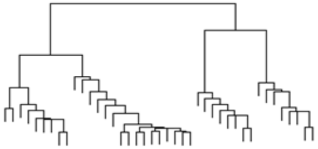
Unsupervised Clustering

P1/P3 LoG Z Uniformity



…

P3/P2 Spatial Correlation



1311 features

2 Prototypical Features

P1/P3 LoG Z Uniformity 12
P1/P3 LoG Z Uniformity 37

P3/P2 Spatial Correlation 8
P3/P2 Spatial Correlation 43

1311 features

Variable Selection (via Lasso)

Logistic Regression Lasso, 10 fold CV

Selected Variables:
Eg. P3/P2 Intensity Mean 3 (var1), P1/P2 DWF D (var5), …., P2 DWF D (var 9)

3 features

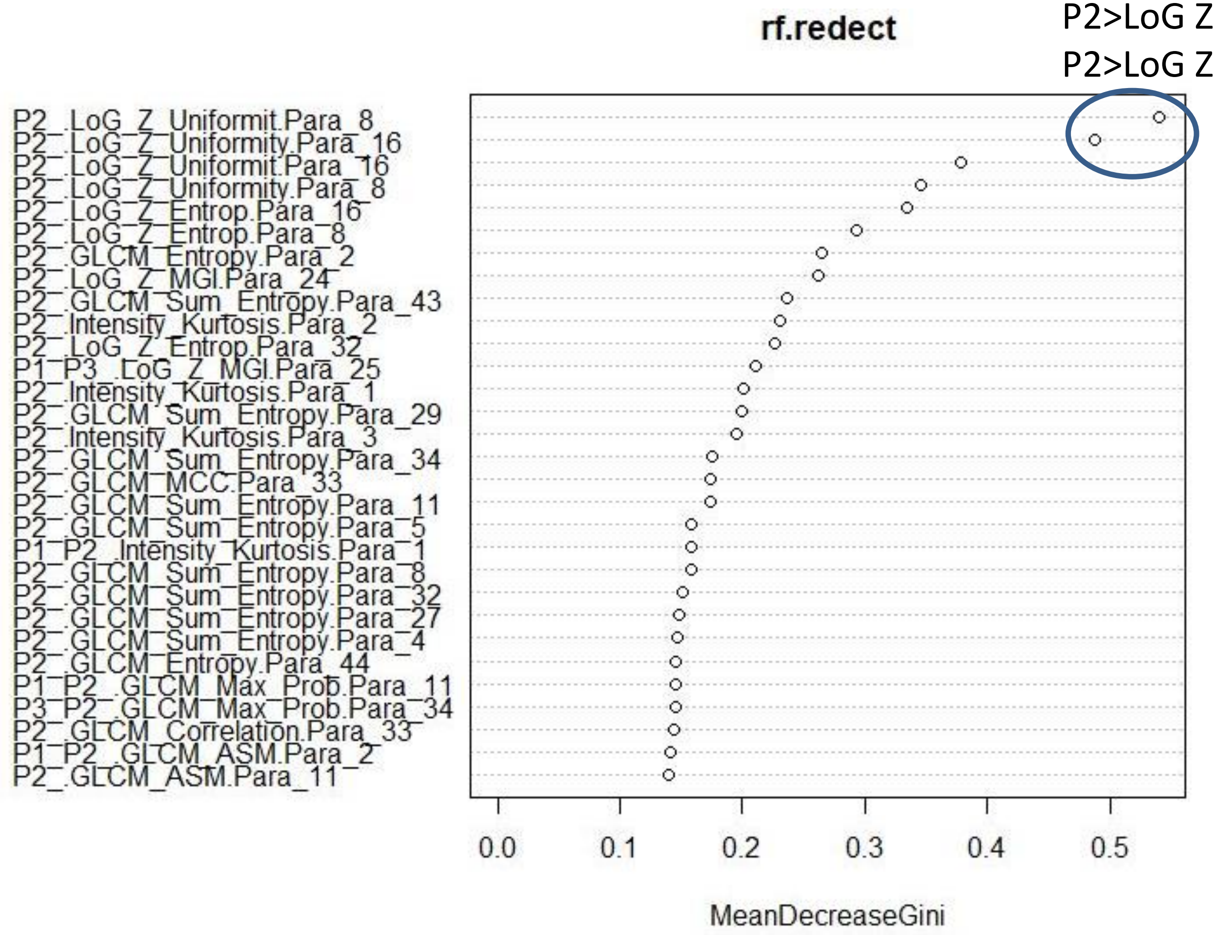
Final Model- Logistic Regression

$$\text{logit}(p) = \beta_o + \beta_1\text{var1} + \beta_2\text{var5} + \ldots + \beta_6\text{var9}$$

Random Forests

190 Observations, 5287 features (same dataset as logistic model)
72 variables tried at each split; 500 trees

P2>LoG Z Uniformit>Para 8	0.540
P2>LoG Z Uniformity>Para 16	0.488



AUC

Lasso Logistic : 0.77

RF: 0.8

Feature space needs a lot of trimming

Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters

Berenguer et al, Radiology 2018; 288:407–415 •

Reproducibility of radiomics for deciphering tumor phenotype with imaging

Binsheng Zhao1,Yongqiang Tan1, Wei-Yann Tsai2, Jing Qi1, Chuanmiao Xie1, Lin Lu1 & Lawrence H. Schwartz[1]

Our data suggest that radiomic features are reproducible over a wide range of imaging settings. However, smooth and sharp reconstruction algorithms should not be used interchangeably. These findings will raise awareness of the importance of properly setting imaging acquisition parameters in radiomics/ radiogenomics research.

Marker evaluation revisited

- **Discovery phase studies:**
  1. **typically based on existing databases**
  2. **assess diagnostic/predictive performance**
  3. **seek to optimize performance**
  4. **need to assess reproducibility of marker results**
- **Central question:**

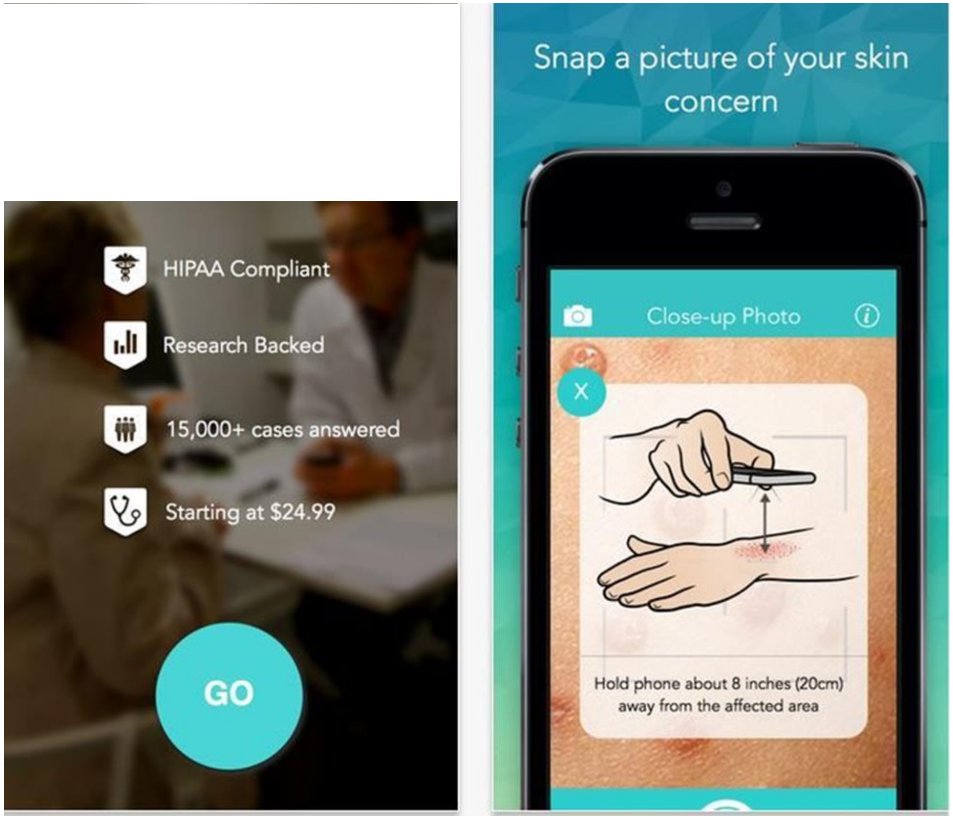
  Is the marker stable, reproducible, and promising enough to move to clinical evaluation?
- **Current radiomics marker research is mainly in the discovery stage.**

Machine learning in apps

Smartphone apps for melanoma detection

- **Large number of apps available.**
- **Technically sophisticated algorithms (e.g. using fractals) for pattern recognition are implemented.**
- **Store and transmit images.**
- **Can compare images taken longitudinally**

Example



Snap a picture of your skin concern

Hold phone about 8 inches (20cm) away from the affected area

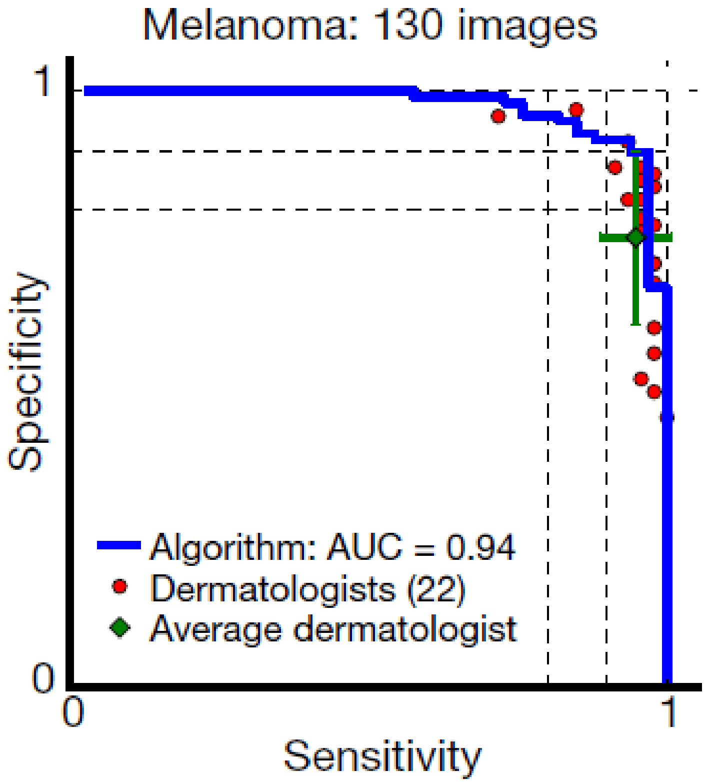
Deep learning potential

Dermatologist-level classification of skin cancer with deep neural networks

Esteva et al, Nature 2017

The CNN achieves performance on par with all tested experts across both tasks, demonstrating an artificial intelligence capable of classifying skin cancer with a level of competence comparable to dermatologists. Outfitted with deep neural networks, mobile devices can potentially extend the reach of dermatologists outside of the clinic. It is projected that 6.3 billion smartphone subscriptions will exist by the year 2021 (ref. 13) and can therefore potentially provide low-cost universal access to vital diagnostic care.

Comparison of accuracies in retrospective reader study



Smartphone-Based Applications for Skin Monitoring and Melanoma Detection

Elizabeth Chao, MD, PhD[a], Chelsea K. Meenan, BS[b],
Laura K. Ferris, MD, PhD[a,*]

Dermatol Clin 35 (2017) 551–557
http://dx.doi.org/10.1016/j.det.2017.06.014

- Despite the abundance of apps …, few have been evaluated for clinical efficacy and none has been sufficiently accurate and reliable using established research methodologies.
- … currently no established quality standards or regulatory oversight of mobile medical apps to ensure patient safety and minimize harm.
- …..important ethical concerns regarding patient confidentiality, informed consent, transparency of data ownership, and data privacy protection.
- Further studies are needed to assess the safety and efficacy ….

Regulating machine learning in devices

FDA approved deep learning software

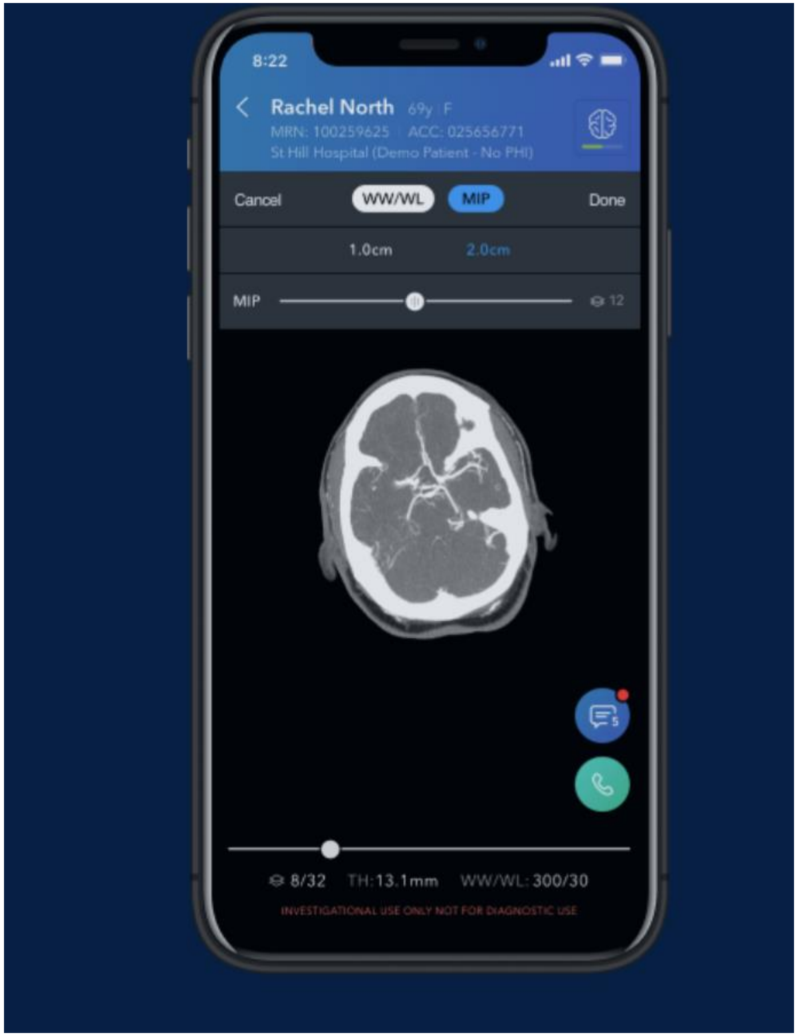
Approved indications for Oncology suite (Jan 2018)

Arterys Oncology DL is a medical diagnostic application for viewing, manipulation, 3D-visualization and comparison of medical images from multiple imaging modalities and/or multiple time-points. The application supports anatomical datasets, such as CT or MR. The images can be viewed in a number of output formats including MIP and volume rendering.

Arterys Oncology DL is designed to support the oncological workflow by helping the user confirm the absence or presence of lesions, including evaluation, quantification, follow-up and documentation of any such lesions.

Note: **The clinician retains the ultimate responsibility for making the pertinent diagnosis** based on their standard practices and visual comparison of the separate unregistered images. Arterys Oncology DL is a **complement** to these standard procedures

FDA approves VizAI clinical decision support

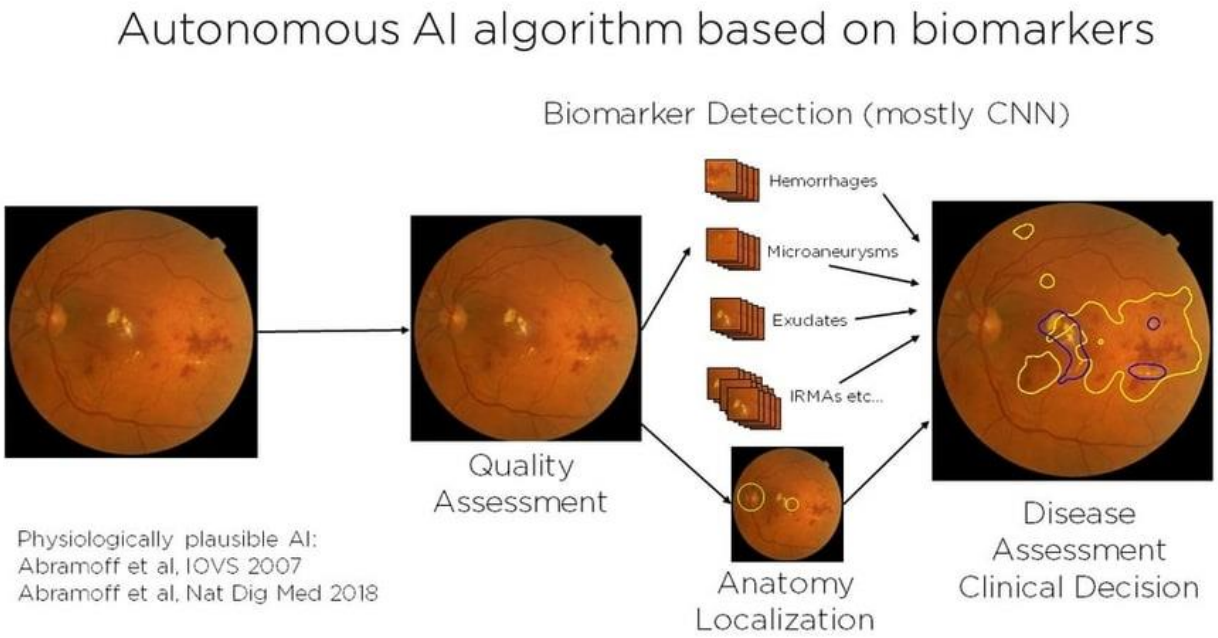


Rachel North 69y | F
MRN: 100259625 ACC: 025656771
St Hill Hospital (Demo Patient - No PHI)

From the FDA press release:

The Viz.AI Contact application is intended to be used by neurovascular specialists, such as vascular neurologists, neuro-interventional specialists or other professionals with similar training. The application is limited to analysis of imaging data and should not be used as a replacement of a full patient evaluation or solely relied upon to make or confirm a diagnosis

FDA approves Idx_DR for dx of diabetic retinopathy



Autonomous AI algorithm based on biomarkers

Physiologically plausible AI:
Abramoff et al, IOVS 2007
Abramoff et al, Nat Dig Med 2018

From FDA press release:

IDx-DR is the first device authorized for marketing that provides a <u>screening decision</u> without the need for a clinician to also interpret the image or results, which makes it usable by health care providers who may not normally be involved in eye care.

DL and Radiomics regulated as CAD

- Parsimonious solution, for now.
- Increasing reliance on CAD likely.
- Reliability and safety of need to be assessed,
- Especially of DL:
  - Face validity of results?
  - Long term properties of algorithms?
  - Under what conditions is performance guaranteed to meet minimum standards?

Commentary

- An avalanche of markers: *Many potential markers. How to prioritize for clinical studies?*
- Software/modalities evolves rapidly: *Moving target: When should evaluation take place?*
- Variability: *by machine, patient cohort*
- Reproducibility: *needs to be established*
- *Appropriate training, calibration*
- *Performance is not guaranteed. Safety and performance monitoring*
- *Face validity of results lacking.*

Collaborators

Samantha Morrison, AM

Jon Steingrimsson, PhD

Thank you!